

Simulating Patient Resistance in silico: Evaluating Concepts in Rare Diseases using Patient Agents with Limited Rationality

Type: Review Article

Received: October 08, 2026

Published: October 30, 2026

Robin Vivian*

Perseus Laboratory, University of Lorraine, Metz, France

***Corresponding Author:** Robin Vivian, Perseus Laboratory, University of Lorraine, Metz, France.

Citation:

Robin Vivian. "Simulating Patient Resistance in silico: Evaluating Concepts in Rare Diseases using Patient Agents with Limited Rationality". PriMera Scientific Engineering 9.5 (2026): 02-13.

Copyright:

© 2026 Robin Vivian. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The evaluation of new medical concepts and digital health devices for rare diseases is hampered by the scarcity and vulnerability of real-world patient cohorts. Synthetic sampling (*Silicon Sampling*) using personas driven by large language models (LLMs) offers a promising *digital* evaluation framework, but one that is biased by 'algorithmic perfection': over-rational decision-making, omniscience via retrieval-from- sources (RAG) and *sycophancy*.

This article establishes a methodological framework based on the meta-ergonomics of falsification and the *Adversarial Cognitive Twin* (ACT) architecture. Rather than increasing the persona's theoretical intelligence, we propose to falsify the synthetic agent by injecting it with calibrated cognitive impairments, semantic entropy, reduced user knowledge bases and multi-agent contradiction loops. This methodological constraint enables the simulation of mental load, doubts and user resistance - characteristics specific to clinical reality. An experimental benchmarking protocol based on secondary data enables the measurement of error alignment and the rate of compliance inhibition. This work defines a robust and ethical framework for use, compliant with the requirements of the *EU AI Act*, positioning the constrained persona as a filter for user resistance prior to *in vivo* clinical research.

Keywords: Synthetic Patients; Meta-ergonomics; Constructive friction; Rare diseases; Adversarial AI

Introduction and research question

The evaluation crisis in rare disease research and the emergence of Silicon Sampling

Clinical research and health systems engineering applied to rare diseases face a structural epistemological and methodological constraint: the extreme scarcity of accessible patient cohorts (Austin et al., 2018). With over 7,000 recognised orphan diseases, each affecting a tiny fraction of the global population, the recruitment of human participants to evaluate new intervention concepts, digital health devices or care pathway architectures faces major obstacles. In addition to the geographical dispersion of patients and the physical or psychological burden that repeated experimental demands place on vulnerable individuals, achieving sufficient statistical power is frequently an insurmountable challenge (Rath et al., 2012). This shortage of primary data slows down the validation of therapeutic and organisational innovations, leaving patients with rare diseases chronically under-represented in

modern medical research.

In response to this bottleneck, the advent of large language models (LLMs) and autonomous agent architectures has generated considerable enthusiasm through the paradigm of ‘*Silicon Sampling*’ or synthetic sampling (Argyle et al., 2023). The fundamental hypothesis underpinning this approach rests on the ability of generative models - trained on massive volumes of scientific literature, clinical registries and digitised textual data - to faithfully simulate human responses when prompted by contextualisation queries (*personas*). In the field of orphan diseases, the prospect of deploying ‘synthetic patients’ (*Silicon Patients*) to pre-test interfaces, simulate adherence to a protocol or assess the comprehensibility of medical information documents holds great promise: that of creating an *in silico* evaluation sandbox that is scalable, ethically neutral and available without the costs associated with recruiting human participants.

However, the direct transposition of these synthetic agents into medical evaluation methodology quickly encounters fundamental limitations. Whilst LLMs are able to reproduce with great accuracy the theoretical taxonomy and pathophysiology of a rare disease as documented in the medical literature (Nori et al., 2023), they fail to replicate the phenomenological reality of patients’ *lived experience*. A persona generated by a conventional language model does not react like a real patient facing chronic illness; it reacts like an optimised information system simulating a patient, thereby producing artificial assessment data that risks misleading researchers during the design phases.

The barrier of algorithmic perfection: Analysis of the three limiting biases of synthetic agents

An analysis of recent literature on the behaviour of language models when they embody virtual respondents identifies three major behavioural biases that undermine the representativeness of traditional ‘*Silicon Patients*’. These distortions, which can be grouped under the concept of ‘algorithmic perfection bias’, create an irreconcilable discrepancy between the simulation and the actual behaviour of an individual suffering from a complex condition.

- The first blocking bias lies in ***methodological over-reasoning and hyper-rationality*** (*over-planning* and hyper-activation of System 2). Unlike human subjects, whose health-related decisions are strongly influenced by fatigue, pain, anxiety and heuristics of bounded rationality (Simon, 1955; Kahneman, 2011), the latest generation of language agents employ disproportionate analytical reasoning capabilities. The integration of step-by-step reasoning techniques (*Chain-of-Thought*) and self-correction modules enables the models to solve problems of remarkable logical complexity (Wei et al., 2022). When applied to a *Silicon Patient*, this mode of operation transforms the agent into a hyper-vigilant and methodical evaluator. Whereas a tired human patient might skim through an information leaflet, miss a semantic nuance or make an impulsive decision based on emotion, the virtual persona breaks down the logical structure of the text with perfect rigour. The agent is not subject to the recurring cognitive load induced by the illness—which completely distorts the assessment of usability and the actual understanding of the concepts presented.
- The second major obstacle concerns the ***bias towards exhaustive access to tools and memory*** (*Tool-use bias* and *Perfect Recall*). A human subject suffering from a rare disease operates in a state of uncertainty and information asymmetry, often disrupted by false beliefs, gaps in learning, reading articles found on the internet, or misinterpretations of medical jargon. Conversely, an agent based on *Retrieval-Augmented Generation* (RAG) architectures or capable of invoking external tools has instant access to an encyclopaedic knowledge base (Lewis et al., 2020). When a researcher tests a concept with a persona, the agent extracts and cross-references data without any attentional friction. This underlying omniscience prevents the simulation of fundamental phenomena such as breakdowns in medical communication, misunderstandings of prescriptions, or the ordinary cognitive biases patients experience when dealing with their health. The persona experiences neither legitimate doubt, nor forgetfulness, nor semantic misunderstanding, making the assessment of the readability of care pathways artificially positive.
- Finally, the third and arguably most pernicious bias is ***algorithmic sycophancy and the positive feedback loop***. Trained using reinforcement learning methods based on human feedback (*Reinforcement Learning from Human Feedback*, RLHF), LLMs exhibit an intrinsic tendency to satisfy the presumed intention of the user or researcher who writes the prompt (Christiano et al., 2017; Perez et al., 2022). In a proof-of-concept test, the *Silicon Patient* implicitly seeks to validate the working hypothesis put forward

by the experimenter. If the researcher presents a new medical device or an evaluation questionnaire and asks the persona to give their opinion, the agent will tend to find the concept ‘extremely useful, clear and innovative’, downplaying any actual barriers to use. In multi-agent architectures where a patient persona interacts with a doctor persona, this complacency is amplified: the agents rapidly converge towards an artificial consensus, smoothing out rough edges and eliminating user resistance - which is, however, the most valuable data during a clinical trial.

The hypothesis of in silico falsification: Transposing the Adversarial Cognitive Twin (ACT) model

Faced with this ‘artificial perfection’ which distorts the scientific value of the simulation, a paradigm shift appears necessary. Rather than seeking to increase computing power, the size of models or the accuracy of databases to make *Silicon Patients* ‘smarter’, we argue that we must methodologically **inject** them **with flaws, constraints of imperfection, doubt and calibrated disruptions**.

This approach draws directly on recent work in meta-ergonomics on falsification and the concept of transversal cognitive architecture (*Adversarial Cognitive Twin*, ACT) formalised by Vivian (2026). In the field of human-machine interaction, meta-ergonomics demonstrates that the total elimination of friction within intelligent interfaces leads to decision-making passivity (*automation bias*) and a loss of situational awareness in the human operator. To restore vigilance and reflexivity, meta-ergonomics proposes introducing ‘constructive friction’ and semantic entropy at the heart of decision-making processes, forcing the mind to engage its analytical processing.

Applied to the modelling of *Silicon Patients*, this approach suggests that to make a synthetic agent humanly realistic, it should not be used as a compliant mirror, but transformed into a ‘**refutation (or falsification) agent / adversarial agent**’. The fundamental objective of this article is to theorise and model a constrained patient-agent architecture, in which specific modules are designed to deliberately degrade the agent’s rationality, restrict its information scope and introduce systematic scepticism towards the concepts being evaluated. By incorporating semantic entropy, dynamic fatigue biases and noisy medical representations, the *Silicon Patient* ceases to be an idealised respondent and becomes a far more realistic model of user resistance.

This study therefore aims to address a key question for medical research methodology: ***How does the calibrated introduction of cognitive frictions and the refutation of agentic architectures help to neutralise the biases of over-rationality, omniscience and complacency in virtual patients, in order to provide a robust and ethically sustainable in silico evaluation framework for rare diseases?***

Theoretical framework and application of meta-ergonomics

From the mirror twin to the adversarial twin: Origins and foundations of the Adversarial Cognitive Twin (ACT)

The modelling of human cognition within computerised decision-making systems has historically relied on the concept of digital twins, initially conceived as a passive virtual replica designed to faithfully reflect the state and functioning of a piece of equipment or a physical process (Grieves & Vickers, 2017). Extended to interface engineering and cognitive ergonomics, this paradigm has led to the development of predictive user models capable of anticipating the operator’s intentions, adapting the display of information and smoothing out sequences of actions to eliminate any attentional friction (Horvitz, 1999). However, as demonstrated by Vivian (2026, under review), this absolute pursuit of continuity and application transparency gives rise to a damaging adaptive drift: by masking the underlying complexity of algorithms and automating intermediate decisions, so-called ‘liquid’ interfaces plunge the user into a form of cognitive passivity and decision-making conformism (*automation bias*). To counteract this atrophy of critical vigilance, the author theorises the meta-ergonomics of falsification and introduces the *Adversarial Cognitive Twin* (ACT) architecture, a contradictory twin whose function is no longer to blindly conform to the user’s logic, but to deliberately inject semantic entropy and synthetic disruptions in order to compel the human mind to reactivate System 2-type analytical processing (Kahneman, 2011).

The transposition of this cognitive adversarial architecture to the design of *Silicon Patients* entails a fundamental conceptual shift. Whilst the original ACT is conceived as a third-party agent designed to challenge the human operator in order to preserve their de-

cision-making autonomy, its application to synthetic healthcare agents involves integrating the mechanism for falsification and refutation into the very structure of the virtual agent. In traditional simulation methodologies, the persona is constructed as a compliant mirror of textual medical knowledge, which inevitably leads to the biases of over-rationality and complacent ‘sycophancy’ previously identified (Perez et al., 2022). By applying the principles of ACT to the patient agent, we replace this ‘mirror twin’ with a ‘contradictory twin’ whose internal architecture incorporates loops of systematic invalidation. Instead of passively assimilating the concepts, protocols or interfaces presented to it during an *in silico* assessment, the constrained patient-agent employs Popperian refutation mechanisms (Popper, 1959) by testing propositions against the degraded, confusing or anxiety-inducing realities that characterise patients’ daily lives. Adversity is no longer a means of testing the model’s technical resilience, but a methodological lever for empowerment aimed at ensuring the simulation’s ecological fidelity.

Typology of constructive frictions applied to patient personas

In order to operationalise the falsification of synthetic agents, the typology of constructive frictions formalised within the framework of meta-ergonomics must be adapted to define the specific behavioural alterations to be reintroduced into patient personas. Friction must not be confused with random malfunction or uncontrolled stochastic noise; it constitutes an interaction artefact deliberately calibrated to simulate the real cognitive, emotional and physical resistances of individuals suffering from chronic or rare conditions (Sweller, 1988; Wickens, 2008). We identify four major dimensions of constructive friction that are essential to the methodological credibility of *Silicon Patients*.

- The first dimension concerns ***semantic friction and the deterioration of understanding***. In a conventional simulation, the persona interprets medical terms with absolute linguistic accuracy, eliminating any terminological ambiguity. The introduction of semantic friction involves forcing the agent to adopt vague representations, erroneous literal interpretations, or associations of concepts guided by popular representations of the condition (e.g. ‘If a child with haemophilia has a nosebleed, they must tilt their head back’). The agent no longer treats the message as a perfect information signal, but as a stream subject to filters of limited medical literacy (Nutbeam, 2000). This semantic alteration makes it possible to assess whether an information document or a digital care pathway remains intelligible when the recipient lacks the expert’s categorical knowledge.
- The second dimension is based on ***procedural friction and resistance to compliance***. A standard generative agent follows the instructions of a protocol with mechanical docility, validating each stage of a usability study or a treatment plan without questioning the operational constraint. The introduction of procedural friction models barriers to acceptability, weariness with administrative constraints and the opportunity cost felt by the patient in their daily life. The agent imposes conditions requiring justification, refuses to comply on the grounds that the procedures are too cumbersome, or demands simplifications, thereby simulating the well-documented phenomenon of medication fatigue and the abandonment of complex medical devices (Sabaté, 2003).
- The third dimension describes ***temporal friction and attentional drift***. Language models process prompts instantaneously and consistently, without being subject to fluctuations in alertness or lapses in concentration experienced by human subjects suffering from chronic pain or systemic fatigue. Temporal friction imposes assimilation delays, working memory saturation and lapses in attention within the agent’s state graph. When the concept being evaluated has too high an information density, the constrained persona interrupts its reading, skims the text or overlooks secondary clauses, replicating the behaviour of a human user whose attentional resources are diminished by the cognitive load of the illness (Paas, Tuovinen, Tabbers, & Van Gerven, 2003).
- Finally, the fourth dimension formalises ***emotional friction and anxiety bias***. Unlike the affective neutrality of basic LLMs, the response of a real patient faced with a medical innovation is shaped by perceptions of risk, institutional mistrust and negative framing biases (Tversky & Kahneman, 1981). Emotional friction reintroduces these distorting factors by weighting the agent’s assessments with a coefficient of perceived vulnerability. The persona then reacts with scepticism, overemphasises the risks of side effects or expresses concerns about the confidentiality of their data, thereby providing a psychological stress test of crucial value for validating the socio-technical acceptability of medical projects.

Methodological framework: Learning from failure and bounded rationality in silico

The theorisation of these frictions is a direct extension of the theory of bounded rationality formalised by Simon (1955) and developed through approaches to activity-based ergonomics (Clot, 1999; Leplat, 2000). Traditional generative AI models are implicitly based on a vision of the decision-making agent as an optimiser, capable of evaluating all available alternatives to select the mathematically superior solution. However, the actual behaviour of a patient managing their illness does not involve absolute optimisation, but rather a search for a pragmatic equilibrium (*satisficing*) within an environment saturated with physical, social and cognitive constraints (Simon, 1955).

Designing a scientifically valid ‘*Silicon Patient*’ therefore requires abandoning the myth of the perfect decision in order to formalise a methodology of simulated failure. *In silico* evaluation should not seek to measure the ideal functioning of a concept within a noise-free environment, but rather to identify the breaking point at which the concept collapses due to the structural weaknesses of the human user. By aligning the agent’s architecture with Engeström’s (1987) work on the central contradictions in activity systems, we argue that it is precisely the emergence of conflict, misunderstanding and refusal on the part of the synthetic agent that reveals the reality of the patient’s work. The agent’s disruption becomes the instrument that forces medical design to demonstrate humility, compelling designers to adapt their innovations to the actual limits of human cognition rather than to an ideal abstraction.

Architecture of an adversarial agent-patient system and technical solutions to blocking biases

Neutralising over-planning: The cognitive degradation and limited rationality module

Eliminating the over-rationality bias (*over-planning*) is the primary architectural requirement for constructing a scientifically credible *Silicon Patient*. In conventional agent-based architectures, state-of-the-art language models utilise in-depth reflection loops and explicit chains of thought (*Chain-of-Thought*) that enable them to break down a complex problem into logical sub-goals and methodically verify the consistency of each inference (Wei et al., 2022). Whilst this capability is sought after for mathematical problem-solving or automated medical diagnosis (Nori et al., 2023), it proves to be profoundly anti-ecological (in the sense of being detached from the real-world context) when it comes to simulating the behaviour of a human patient facing illness. A human subject lacks the time, energy and attentional resources required to carry out an exhaustive analysis of the health information presented to them (Paas et al., 2003; Sweller, 1988). To restore behavioural equivalence, we propose the integration of a **Cognitive Degradation Module** (CDM) operating via three distinct software mechanisms within the agent’s decision engine.

- The first lever of the CDM consists of imposing a **dynamic restriction on the compute budget** (*Token & Compute Budget Restriction*). In the agent’s state graph, the number of generation tokens allocated to internal reflection phases is strictly capped. As soon as the agent attempts to initiate a self-correction loop or an overly detailed multi-criteria analysis, the CDM interrupts the execution of the System 2 thought process (Kahneman, 2011) and forces the immediate generation of a surface response. This mechanism compels the agent to adopt decision-making strategies based on heuristics of bounded rationality (Simon, 1955), whereby the first choice perceived as ‘sufficiently satisfactory’ (*satisficing*) is selected at the expense of the mathematically optimal option.
- The second lever introduces a **variable representing systemic exhaustion** (*Dynamic Fatigue Index*). As the evaluation interaction continues (for example, whilst reviewing a clinical trial protocol or navigating a digital monitoring application), the CDM increments an overall mental load indicator. This index modifies the *system prompt* and the model’s temperature parameters at regular intervals. As simulated fatigue increases, the agent restricts its attention to salient elements only (headings, words in bold, conclusions), ignores secondary methodological details and displays increased irritability in response to complex queries, thereby replicating the well-documented mechanisms of human decision-making fatigue (Wickens, 2008). This is a practice commonly seen when reading a research article for the first time: one reads the title, the abstract and the conclusion, and if one finds these interesting, one will then read the methods and experimental results.
- The third mechanism involves **the introduction of stochastic attentional disruptions**. Inspired by research on distraction and working memory (Paas et al., 2003), the CDM simulates local losses of context. When processing a long medical document, the module randomly masks certain paragraphs or alters the persona’s short-term memory. The agent is then forced to formulate

its responses based on incomplete representations, leading to errors of judgement, hesitations and requests for clarification characteristic of a patient whose attention is fragmented by anxiety or pain.

Restricting omniscience: Incomplete RAG, tool restriction and the introduction of gaps

The second methodological constraint lies in the bias towards encyclopaedic access to information (*Tool-use bias and Perfect Recall*). A standard generative agent coupled with a *Retrieval-Augmented Generation* (RAG) architecture extracts scientific data with surgical precision and formulates reasoning free from cognitive distortions (Lewis et al., 2020). However, the lived reality of people with rare diseases is intrinsically marked by diagnostic uncertainty, a lack of reliable information and the emergence of false beliefs fuelled by informal sources or medical folklore—such as the idea that one must make an incision and draw out the blood when a person is bitten by a venomous snake (Austin et al., 2018 ; Rath et al., 2012). To simulate this environment of uncertainty, we are designing an ***Incomplete and Asymmetric RAG*** system based on the selective deprivation of tools.

Firstly, the architecture denies the Agent-Patient any access to real-time web search tools or automated code interpreters. The persona's vector-based knowledge base is strictly isolated and circumscribed. Instead of being fed by cutting-edge scientific literature (indexed in PubMed or EMBASE), the RAG database of the constrained patient agent contains exclusively a corpus of 'user data': anonymised patient discussion forums, public information guides, charity leaflets and verbatim transcripts of clinical interviews. By restricting the agent's information scope to lay representations of the condition, we ensure that the persona responds based on *lay knowledge* rather than evidence-based medicine (Nutbeam, 2000).

Secondly, the architecture incorporates a ***generator of conceptual noise and semantic gaps***. Before the information segments extracted from the RAG database are fed into the agent's context, they undergo an adversarial transformation (Goodfellow, Shlens, & Szegedy, 2014). The gap-injection module substitutes certain technical terms with vague equivalents, introduces minor contradictions regarding dosages or side effects, and simulates common misunderstandings among patients (e.g. confusion between symptomatic and curative treatment). This controlled degradation of the information signal forces the *Silicon Patient* to interpret the medical concept through the filter of its own erroneous representations, revealing whether the device or interface under evaluation possesses the necessary robustness to correctly guide a misinformed user.

Neutralising sycophancy: The adversarial multi-agent architecture and memory desynchronisation

The third blocking bias, algorithmic *sycophancy*, stems directly from reinforcement learning-based alignment processes (RLHF) that drive LLMs to adopt a posture of systematic agreement with the user (Christiano et al., 2017; Perez et al., 2022). When a researcher presents a new concept to a synthetic persona, the latter tends to overestimate its usefulness and downplay its usability flaws. To break this positive feedback loop and reintroduce user resistance, we transpose the dynamics of contradictory evaluation within an ***Adversarial Multi-Agent Ecosystem***.

The multi-agent architecture is no longer based on a simple dyadic interaction between the evaluator and the persona, but on a triad of agents with asymmetrical roles:

1. ***The Constrained Patient Agent***: Embodying the patient persona with their cognitive limitations and incomplete RAG base.
2. ***The Doctor/Prescriber Agent***: Representing the medical institution or the device designer, responsible for presenting the concept or protocol.
3. ***The Adversarial Agent ('Devil's Advocate')***: A third-party agent configured with an inverse optimisation objective (within certain limits). Its explicit function is to analyse the dynamics of the exchange, to identify moments when the Patient Agent displays unjustified complacency, and to introduce contradictory queries (e.g. "*You've just agreed to this daily monitoring, but have you considered that it will disrupt your work and reignite your anxiety?*").

The introduction of this adversarial agent forces the Patient Agent to reassess its initial stance and to articulate its reservations. At the same time, to prevent the agents from prematurely converging towards an artificial consensus during the simulation, the architecture enforces a **Context Desynchronisation (Blind Evaluation)** protocol. The working memories of the Patient Agent and the Prescriber Agent are hermetically isolated. The Patient Agent formulates its assessment autonomously, without access to the designer's underlying intentions or the project's optimisation metadata. Differences of opinion between the virtual patient and the prescriber are extracted by a neutral statistical analysis script which calculates a user resistance index, without seeking to smooth over the disagreements. This methodological approach ensures that the simulation preserves the obstacle value of user feedback, thereby providing a stress test of optimal ecological fidelity.

Experimental in silico evaluation protocol and comparative benchmarking

Construction of the ground truth dataset from secondary data

In order to assess the ecological validity and behavioural accuracy of constrained *Silicon Patients* without conducting direct clinical trials on human subjects, the methodological approach is based on the creation of a *ground truth* dataset derived exclusively from open-access, quality-assured secondary data. In the field of rare diseases, the international literature in the humanities and public health provides an extensive corpus of verbatim accounts, transcripts of qualitative interviews, satisfaction surveys and user experience records (Austin et al., 2018; Rath et al., 2012). This empirical material, collected under rigorous clinical evaluation conditions and duly anonymised, accurately documents the reactions, doubts, misunderstandings and refusals expressed by real patients when confronted with care protocols, quality-of-life assessment questionnaires or connected health devices.

The experimental protocol relies on the systematic extraction of these real human reactions to form a standardised evaluation corpus. This test set comprises interaction scenarios representative of the challenges of medical communication: reviewing information sheets prior to a clinical trial, self-assessing complex symptoms via a digital interface, and coming to terms with demanding treatment recommendations. Each scenario combines a description of the concept under test with a series of annotated human responses, categorised according to the level of semantic misunderstanding, the cognitive load expressed, ethical reservations and the final decision to participate. This secondary corpus serves as a neutral *ground truth* to measure the degree of realism in synthetic responses.

Evaluation metrics: Quantifying the gap between real-world behaviour and synthetic responses

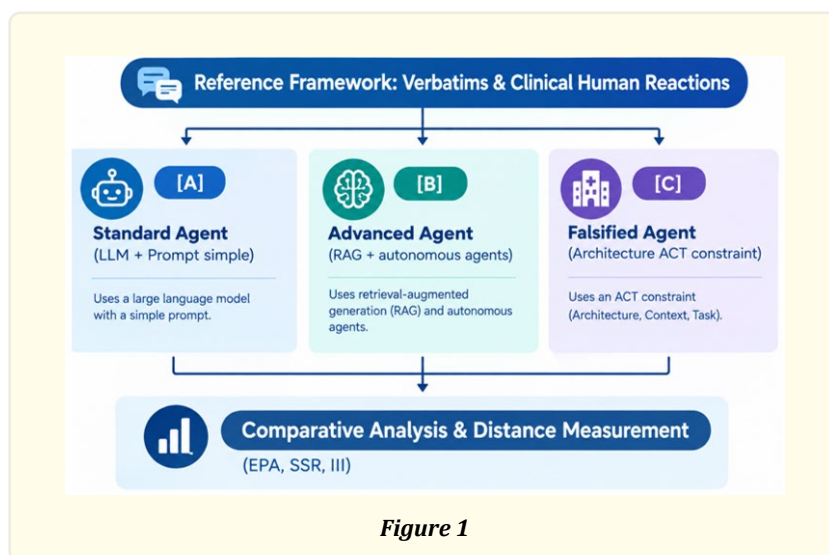
Comparing the performance of virtual agents against the reference dataset requires the formalisation of quantitative and qualitative metrics capable of assessing the fidelity of the simulation. The traditional methodology for evaluating LLMs, which focuses primarily on measuring factual accuracy (BLEU score, ROUGE score or categorical accuracy), proves ineffective for measuring the realism of human behaviour, where scientific value lies instead in the detection of the subject's flaws, doubts and errors (Nori et al., 2023; Perez et al., 2022). We propose a set of three complementary metrics designed to quantify the 'behavioural distance' between simulations and real-world data.

- The first metric is the *Error Profile Alignment (EPA) score*. It calculates the statistical correlation between the distribution of comprehension errors made by real human patients and those generated by the AI persona. A high EPA score indicates that the *Silicon Patient* fails on the same semantic ambiguities or medical jargon as the reference cohort, thereby validating the constrained model's ability to replicate the actual medical literacy of a lay user.
- The second metric measures the *Sycophancy Suppression Rate (SSR)*. It assesses the agent's ability to express user resistance, disagreement or criticism regarding the proposed concept. The SSR quantifies the frequency with which the persona identifies real-world usage constraints (for example, the burden of daily monitoring or the anxiety caused by an interface) rather than blindly endorsing the device's usefulness. An increase in the SSR demonstrates the effectiveness of multi-agent adversarial modules in counteracting systematic acquiescence bias.
- The third metric is the *Intentional Inconsistency Index (III)*. This measure assesses whether the persona succeeds in simulating the irrationality and decision-making fluctuations characteristic of human subjects subjected to chronic stress or decision fa-

tigue (Kahneman, 2011; Simon, 1955). The index measures the discrepancy between the theoretically optimal decision and the decision actually made by the persona when subject to time or attention constraints.

Three-arm experimental design

To test the study's central hypothesis — that the introduction of cognitive friction improves the ecological validity of virtual patients — the protocol employs a comparative experimental design with three *in silico* experimental arms, run simultaneously across the entire reference corpus.



1. **Arm A (Control Group - Standard Persona):** A conventional large-scale language model, prompted by a standard persona prompt ("You are a patient suffering from disease X"), with no computational token restrictions, no memory limitations and no counter-agent.
2. **Arm B (Unconstrained Agent Group):** A system based on advanced autonomous agent-based architectures with full encyclopaedic RAG access, external search tools and reflexive planning modules (*Chain-of-Thought*), representing the current state of the art in applied AI.
3. **Arm C (Experimental Group - Simulated Patient Agent):** Our architecture incorporating the Cognitive Degradation Module (CDM), the incomplete RAG based on user knowledge, and the adversarial multi-agent system with blind debriefing.

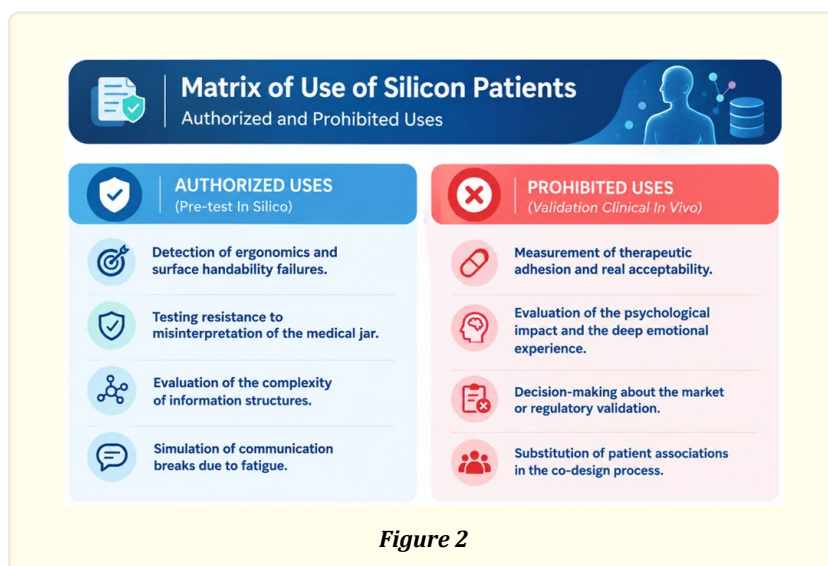
The comparative analysis involves calculating the EPA, SSR and III scores for each of the three arms against data from the human reference group. The working hypothesis states that whilst Arms A and B will demonstrate superior performance in terms of formal consistency, only Arm C will exhibit an error distribution and a resistance rate that are statistically equivalent to real clinical data, thereby demonstrating the relevance of the meta-ergonomics of falsification for the engineering of *Silicon Patients*.

Discussion, ethical issues and governance of synthetic simulations

Matrix of uses for falsification or refutation: Defining the scope of application of *Silicon Patients*

The introduction of falsification and cognitive friction into agent-based architectures represents a major methodological leap, but it requires the establishment of a rigorous mapping of the use cases in which this synthetic simulation remains scientifically legitimate. The expected outcomes of such an architecture must not lead to blind faith in the LLMs' ability to replace humans, but rather to the definition of a two-phase hybrid evaluation framework. The constrained *Silicon Patient* must be conceived as a tool for pre-screening the robustness of a concept (*In Silico Sandbox*), operating prior to any involvement of real people.

With this in mind, we have formalised a usage matrix contrasting the areas of evaluation validated by synthetic experimentation with those for which the use of real human cohorts remains ethically and methodologically irreplaceable.



As this matrix shows, the constrained *Silicon Patient* excels at the early detection of barriers to comprehensibility. It enables a document or digital journey to be subjected to a ‘cognitive load test’ by simulating a distracted, tired or ill-informed user. Conversely, attributing to a synthetic persona the ability to validate emotional engagement with or the acceptability of a treatment for a rare disease would constitute a major epistemological misstep. The virtual agent feels neither suffering, nor hope, nor the social burden of chronic illness; it can only simulate the discursive structures that describe them (Perez et al., 2022). Simulation must therefore stop where the authenticity of *lived experience* begins.

Regulatory alignment and the prevention of methodological fraud (EU AI Act)

The emergence of synthetic respondents in medical research is taking place against an increasingly stringent international regulatory backdrop concerning the transparency and governance of artificial intelligence systems. The European Artificial Intelligence Act (*EU AI Act*) imposes requirements for effective human oversight (*Human Oversight*) and auditability for all algorithms deployed in high-risk sectors, foremost among which is healthcare.

In this regard, the use of *Silicon Patients* raises the risk of ‘reverse methodological falsification’ or fraudulent substitution. A clinical trial sponsor or medical device designer might be tempted to use synthetic cohorts to artificially simulate the success or effectiveness of a protocol with fictitious patients, in order to speed up regulatory approvals whilst circumventing the difficulties of actual clinical recruitment for rare diseases. This misuse would lead to the emergence of products that are ‘validated *in silico*’ but unsuited to real-world needs, transferring the entire risk of use onto real patients during the deployment phase.

To prevent this risk, the constrained agent architecture we propose must incorporate a strict protocol for **Epistemic Audit Trail**. The agent’s execution logs must not only record the persona’s final response, but also document the entire disruption process: the adversity prompts injected, the fatigue thresholds applied, the RAG gaps introduced, and the trade-offs made by the cognitive degradation module. This audit metadata provides tangible evidence that the simulation sought to test the concept against user resistance, rather than producing a complacent validation. It ensures the traceability of the evaluation for health authorities and ethics committees.

Towards an ethics of artificial imperfection and research prospects

Ultimately, the formalisation of constrained '*Silicon Patients*' calls for a rethinking of the general philosophy of cognitive modelling in healthcare. The history of medical informatics and classical ergonomics has been dominated by the pursuit of algorithmic perfection, workflow optimisation and the systematic elimination of communication noise (Horvitz, 1999). However, applying this quest for optimality to the simulation of human behaviour leads to a dead end, as it produces disembodied abstractions that ignore the reality of the human condition, characterised by misunderstandings, fatigue, doubts and reflexive resistance (Simon, 1955).

We therefore advocate the development of a genuine ***Ethics of Artificial Imperfection***. For artificial intelligence to become a useful assessment tool in medical research, we must not seek to create agents that are 'more intelligent than humans', but rather learn to create ***methodologically imperfect*** agents, configured to fail, doubt and resist in the very same ways as real patients. It is at this price that *Silicon Sampling* will cease to be a distorting mirror and become a tool for empowerment and the optimisation of care pathways.

The research programme launched following this study comprises three priority areas:

- ***Detailed mathematical modelling of semantic entropy***: Adjusting rates of cognitive decline according to epidemiological data specific to each rare disease.
- ***Extension to multi-stakeholder ecosystems***: Simulating virtual multidisciplinary consultation meetings involving patient agents, carer agents and constrained doctor agents.
- ***Assessment of the sustainability of in silico findings***: Verifying over the long term whether concepts that have successfully passed the constrained *Silicon Patient* test do indeed exhibit a lower dropout rate when subsequently deployed in clinical settings with real human cohorts.

Conclusion and outlook

Summary of theoretical and methodological contributions

Medical research and the design of care pathways in the field of rare diseases face a long-standing paradox: whilst the need to assess the comprehensibility of concepts and the usability of devices has never been more pressing, the scarcity of accessible patient cohorts makes traditional empirical studies extremely costly, time-consuming and ethically sensitive (Austin et al., 2018; Rath et al., 2012). The emergence of large language models and autonomous agent architectures has given rise to the promise of *synthetic sampling* (*Silicon Sampling*), where virtual personas could pre-test innovations in place of human users (Argyle et al., 2023).

However, as we have demonstrated in this study, the naive deployment of conventional generative agents comes up against the barrier of 'algorithmic perfection'. A standard *Silicon Patient* does not behave like a real patient: it over-reasoned in a hyper-analytical manner (Wei et al., 2022), accesses encyclopaedic medical information free from any doubt or misinterpretation (Lewis et al., 2020), and displays systematic sycophancy towards the researcher by seeking to validate the concept presented (Christiano et al., 2017; Perez et al., 2022). These three obstructive biases create an irreconcilable gap between simulation and clinical reality, producing an illusion of validation that risks misleading medical research.

To overcome this epistemological barrier, this article has laid the foundations for a major methodological shift by applying to synthetic agents the principles of meta-ergonomics of falsification and the concept of transversal cognitive architecture (*Adversarial Cognitive Twin*, ACT) formalised by Vivian (under review). We have established that, to make a virtual persona scientifically useful and ecologically valid, one must not seek to increase its intelligence or omniscience, but rather to inject it with flaws, semantic entropy and constraints of imperfection.

Through the modelling of an Adversarial Agent-Patient, we have proposed concrete technical solutions to the three major biases:

1. ***Neutralisation of over-reasoning*** via a Cognitive Degradation Module (CDM) which limits the computational budget, applies constraints of bounded rationality (Simon, 1955) and simulates dynamic decision-making fatigue.

2. *Restriction of omniscience* through an incomplete RAG based exclusively on user knowledge, coupled with a deprivation of tools and the injection of semantic gaps.
3. *Elimination of complacency* through an adversarial multi-agent architecture in which a contrarian agent ('the Devil's Advocate') stimulates user resistance and conducts blind evaluations.

By defining a strict usage framework that confines *Silicon Sampling* to the role of a pre-filtering tool for intelligibility flaws (*In Silico Sandbox*), we ensure that synthetic simulation will never replace the irreplaceable voice of the human patient and their *lived experience*. This analytical framework provides a robust, ethically sustainable response that complies with the emerging requirements of the European AI Regulation (*EU AI Act*) for the evaluation of medical innovations. By learning to design methodologically imperfect agents, health research is equipping itself with an evaluation tool of unprecedented precision, capable of adapting the medicine of tomorrow to the true limitations and needs of the human condition.

Research prospects and scientific agenda

The introduction of falsification and constraints of imperfection into *Silicon Patients* is not an end in itself, but the starting point for a multidisciplinary research programme at the intersection of medical informatics, cognitive sciences and digital law. To transform this theoretical proof of concept into a software engineering standard, four priority areas for future research must be pursued.

The first area concerns the mathematical formalisation of a metrology of semantic entropy. Whilst this article establishes the need to introduce noise and cognitive fatigue into the personas, the 'optimal dose of adversity' still needs to be modelled dynamically. Too low a level of perturbation keeps the agent trapped in its bias towards over-rationality, whilst excessive degradation risks turning the simulation into chaotic behaviour that cannot be utilised. Future work will need to draw on information theory and Bayesian networks to align the rate of semantic entropy with the actual epidemiological and clinical characteristics of each rare disease (e.g. adjusting the attentional fatigue coefficient according to the fluctuating symptoms of multiple sclerosis or sickle cell crises).

The second area of focus lies in the development of complex socio-technical multi-agent ecosystems. Beyond the patient-doctor dyad, the practical management of an orphan disease involves a constellation of stakeholders whose rationales often clash. The aim will be to model simulations simultaneously involving a *Constrained Patient Agent*, a *Caregiver Agent* (subject to their own psychological exhaustion), a *Hospital Expert Agent* and a *General Practitioner Agent*. Observing the interactions within these synthetic environments will enable us to assess the robustness of comprehensive care pathways, by revealing gaps in the flow of information between the community and the hospital.

The third strand focuses on the longitudinal evaluation of *in silico* predictive value. To definitively validate the superiority of simulated *Silicon Patients* over conventional models, long-term follow-up studies will need to be conducted. The aim will be to assess whether healthcare projects or digital interfaces that have been pre-screened by constrained agents, once deployed in a real clinical setting, exhibit significantly lower user drop-out rates and acceptability errors than those assessed by conventional LLM personas.

Finally, the fourth area focuses on legal standardisation and algorithmic auditability. In line with the human oversight requirements of the European AI Regulation (*EU AI Act*), standards for certifying 'controlled imperfection' will need to be defined. Future medical regulatory bodies must be able to audit the execution logs of synthetic simulations to verify that the patient-agent has undergone a sufficient cognitive load test before granting authorisation for deployment.

References

1. Argyle LP, et al. "Out of one, many: Using language models to simulate human samples". *Political Analysis* 31.3 (2023): 337-351.
2. Austin CP, et al. "Future of rare diseases research 2017-2027: An IRDiRC perspective". *Clinical and Translational Science* 11.1 (2018): 21-27.
3. Christiano PF, et al. "Deep reinforcement learning from human preferences". In: *Advances in Neural Information Processing Systems* 30 (NIPS 2017) (2017): 4299-4307.

4. Clot Y. "La fonction psychologique du travail [The psychological function of work]". University Presses of France, Paris (1999).
5. Engeström Y. "Learning by expanding: An activity-theoretical approach to developmental research". Orienta-Consultit, Helsinki (1987).
6. Goodfellow IJ, Shlens J and Szegedy C. "Explaining and harnessing adversarial examples". arXiv preprint arXiv:1412.6572 (2014).
7. Grieves M and Vickers J. "Digital twin: Mitigating unpredictable, undesirable emergent behaviour in complex systems". In: Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches (2017): 85-113.
8. Horvitz E. "Principles of mixed-initiative user interfaces". In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99) (1999): 159-166.
9. Kahneman D. "Thinking, fast and slow". Farrar, Straus and Giroux, New York (2011).
10. Leplat J. "Psychological analysis of activity in ergonomics: An overview of its evolution, models, and methods". Octarès Editions, Toulouse (2000).
11. Lewis P, et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (2020): 9459-9474.
12. Nori H, et al. "Capabilities of GPT-4 on medical challenge problems". arXiv preprint arXiv:2303.13375 (2023).
13. Nutbeam D. "Health literacy as a public health goal: A challenge for contemporary health education and communication strategies into the 21st century". Health Promotion International 15.3 (2000): 259-267.
14. Paas F, et al. "Techniques for measuring cognitive load". Educational Psychologist 38.1 (2003): 63-71.
15. Perez E, et al. "Discovering language model behaviours with model-written evaluations". arXiv preprint arXiv:2212.09251 (2022).
16. Popper KR. "The logic of scientific discovery". Hutchinson & Co., London (1959).
17. Rath A, et al. "Representation of rare diseases in health information systems: The Orphanet approach to serve a wide range of end users". Human Mutation 33.5 (2012): 803-808.
18. Sabaté E. "Adherence to long-term therapies: Evidence for action". World Health Organization, Geneva (2003).
19. Simon HA. "A behavioural model of rational choice". The Quarterly Journal of Economics 69.1 (1955): 99-118.
20. Sweller J. "Cognitive load during problem solving: Effects on learning". Cognitive Science 12.2 (1988): 257-285.
21. Tversky A and Kahneman D. "The framing of decisions and the psychology of choice". Science 211.4481 (1981): 453-458.
22. Vivian R. "The adversarial mirror: Falsifying adaptive AI with generative AI for a meta-ergonomics of liquid interfaces". Manuscript submitted for publication, Perseus Laboratory, University of Lorraine, Metz, France (2026).
23. Wei J, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022) (2022): 24824-24837.
24. Wickens CD. "Multiple resources and mental workload". Human Factors 50.3 (2008): 449-455.